DRAPES: Diffusion for weakly supervised searches

Hammers and Nails, Swiss Edition 2023

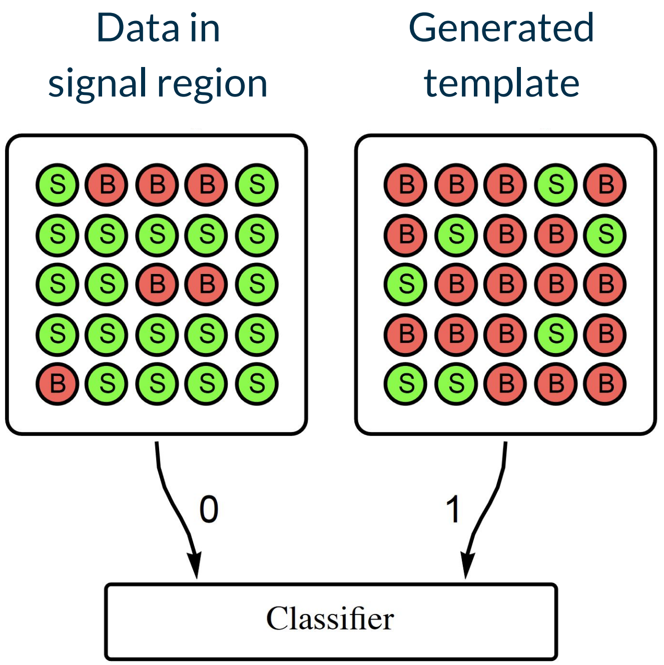
Debajyoti Sengupta, Matthew Leigh, Johnny Raine, Sam Klein, Tobias Golling

UNIVERSITÉ DE GENÈVE

Established Task

- Use CWOLA to look for anomalous samples

- Signal region contains: **B+S**
- Template should contain: **B (+S')**



Weakly Supervised Regime

- Generate a background template in signal region (SR)
- Train a classifier b/w template and SR data.

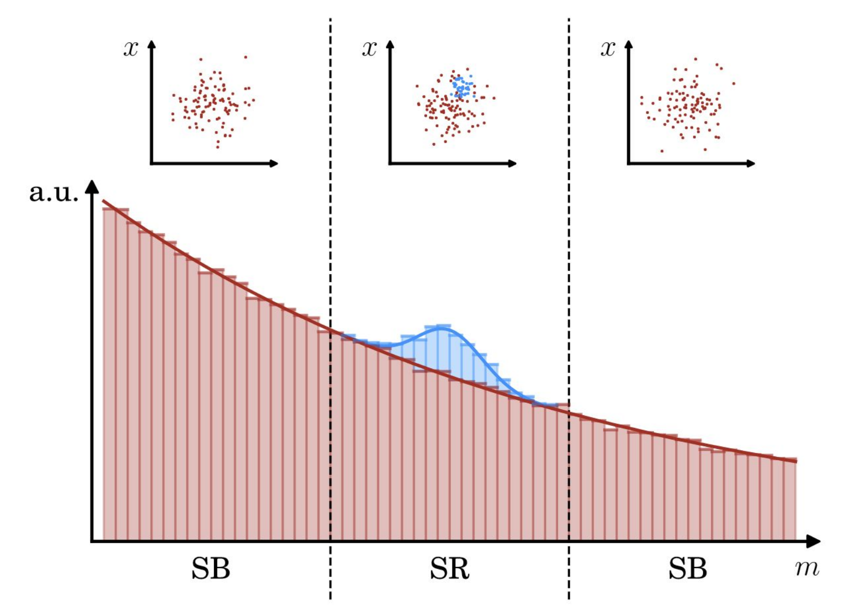
Existing template generators:

1. **CATHODE**: flow based conditional generator (data driven)

2. **CURTAINs**: flow based feature morpher (data driven)

3. **FETA**: flow based feature morpher (simulation assisted)

4. **SALAD**: classifier based reweighting (simulation assisted)



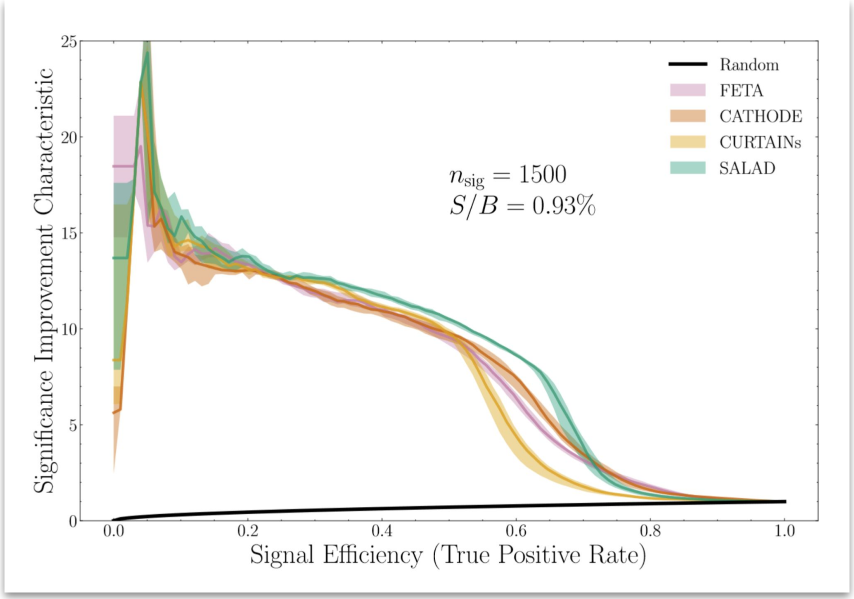
Weakly Supervised Regime

- Results for the LHCO RnD dataset
  - Background: QCD dijets
  - Signal: W' → X(qq) Y(qq)

- Features: $m_{J_1},\ \Delta m_J = m_{J_1} - m_{J_2},\ \tau_{21}^{J_1},\ \tau_{21}^{J_2},\ \Delta R_{JJ} = \sqrt{\Delta\eta^2 + \Delta\phi^2}$

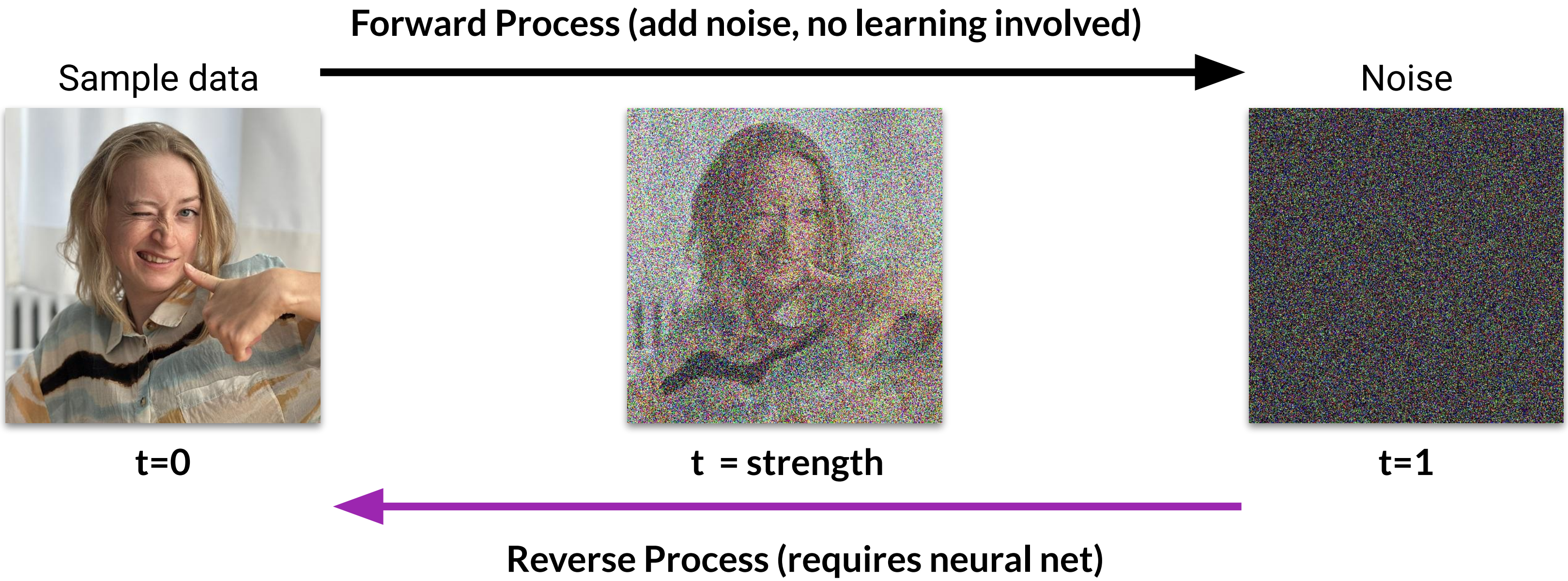
SIC = $\varepsilon_s / \sqrt{\varepsilon_b}$ as a function of ε_s for 1500 signal samples doped in.

All methods perform comparatively well in regions of interest



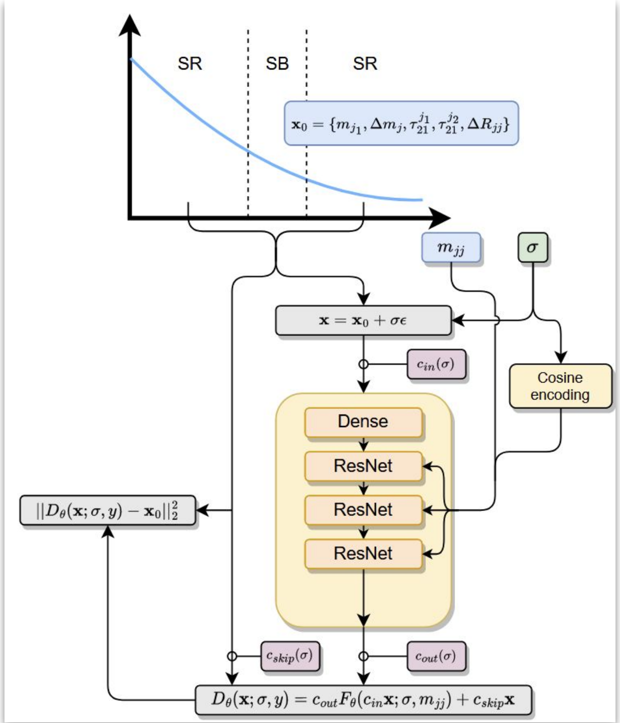
from : 2307.11157

Drapes: Denoising resonant anomalies by perturbing existing samples



Drapes: Training and inference

- ➔ Dense residual network
- ➔ EDM **Diffusion** setup (PC-Droid)
- ➔ Train on **SIDEBAND DATA,** Condition on mass.



- To Generate template:
  - Sample data → add noise → sample mass → denoise

Considerations:

1. Where the data is sampled from
2. How much noise is added and then denoised

Drapes Variants: Where is the data sampled from?

DRAPES SB:

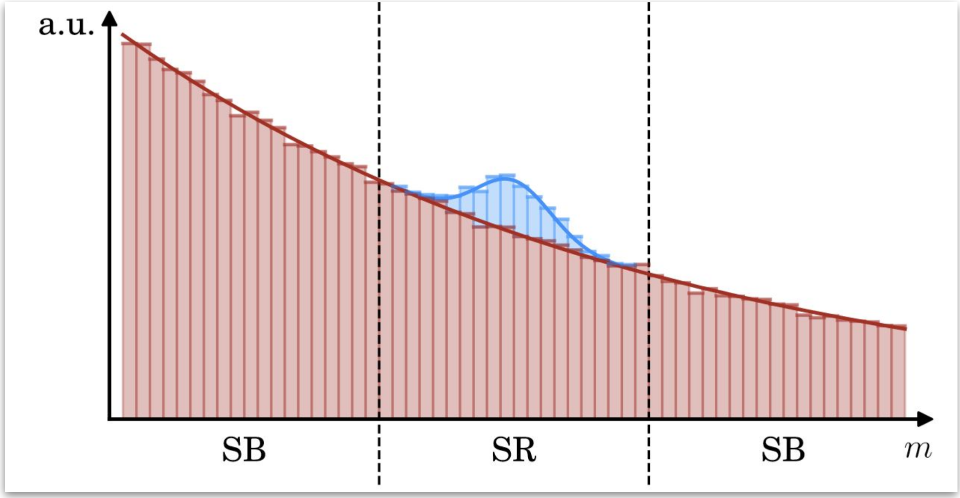
- Sample data (t = 0) from SB
- Give sample new mass (CURTAINs mode)

DRAPES SR:

- Sample data (t=0) from SR

DRAPES MC:

- Sample data (t=0) from MC (FETA mode)



Drapes Variants: How much noise is added and denoised?

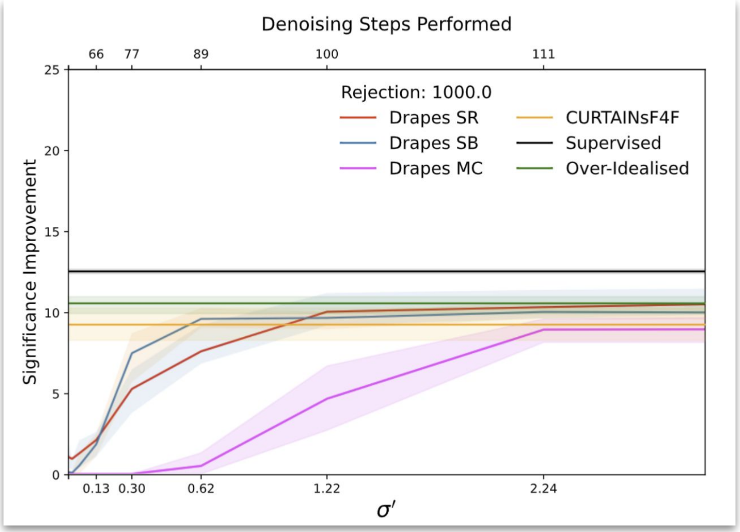
Drapes Φ :

- Full noise and denoise (corresponding to a Gaussian of width 80, and back) (CATHODE Mode)

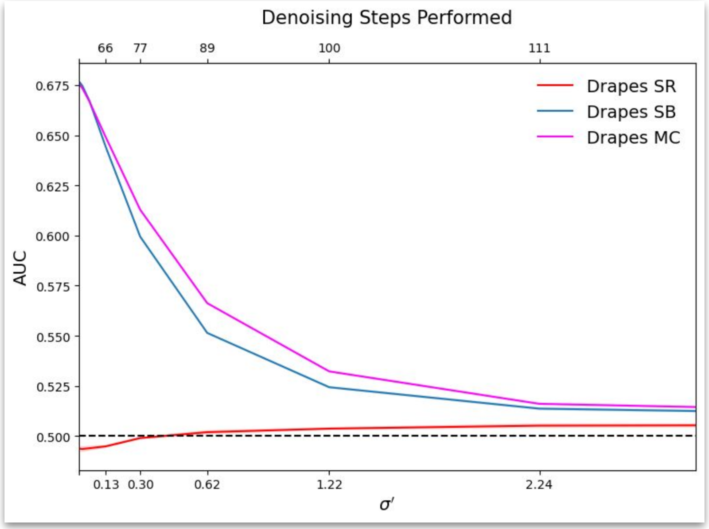
Can also choose to only add a fraction of the noise, and then denoise.

- i.e. instead of sigma = 80, stop at sigma = sigma’ and denoise.
- Performing fewer diffusion steps ⇒ faster

Effect of partial diffusion



SIC(1E3) as a function of sigma’



AUC for template vs background as a function of sigma’

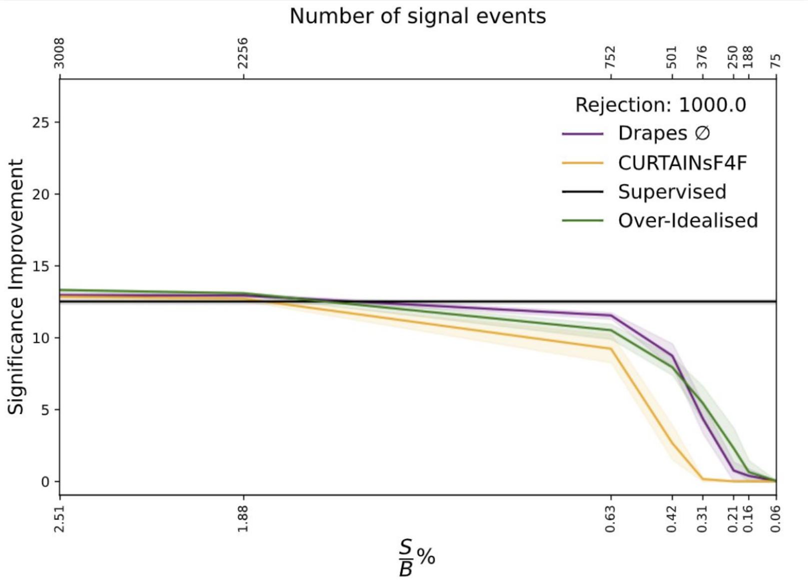
Performance ~ saturates at sigma’ = 2.24, also where AUCs <= 0.52:

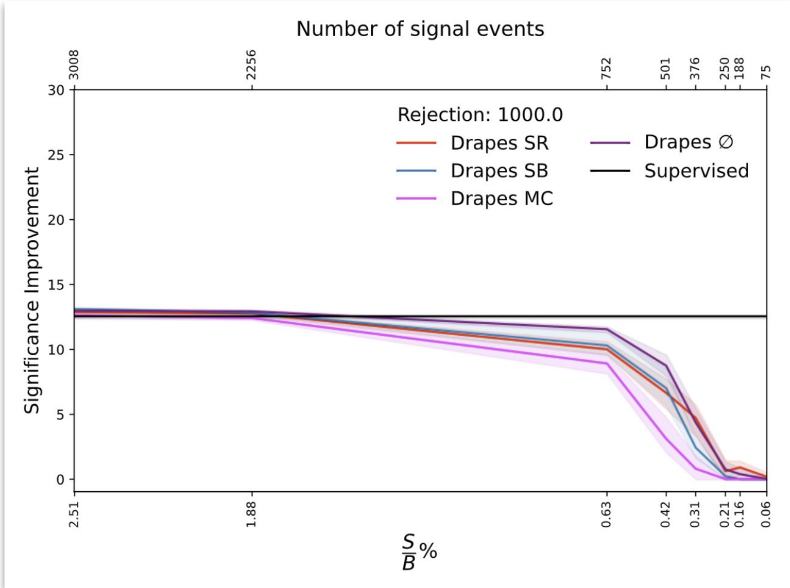
→ Good template reconstruction + good performance

→ Not full denoising → saves on time.

Performance

Performance as a function of signal present



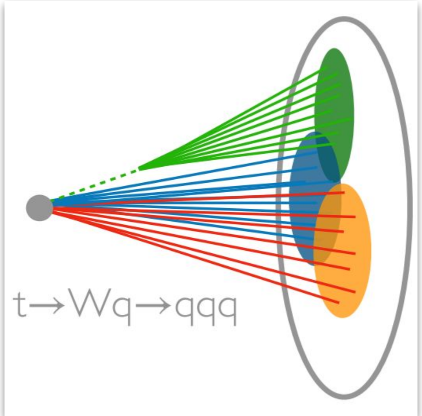


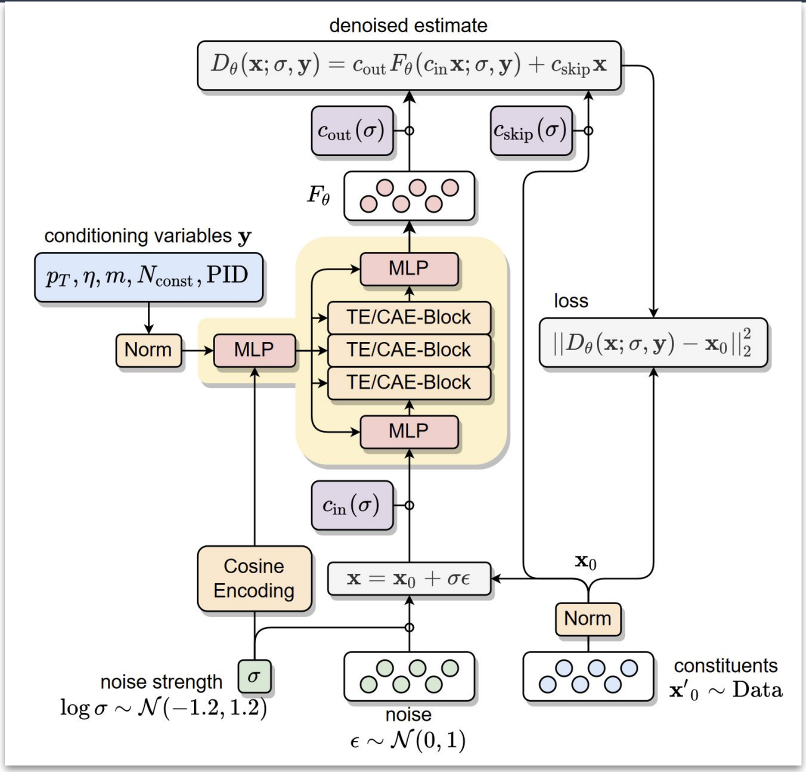
Drapes Φ outperforms existing competition across a wide range!

Drapes for constituent level

Instead of the high level features, train a diffusion model to generate the jet point cloud.

Use Droid model to conditionally generate jets

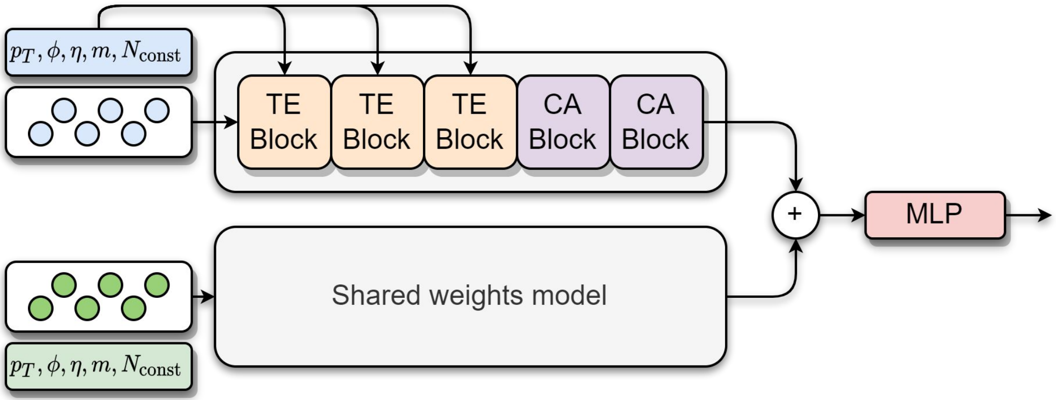




https://arxiv.org/abs/2307.06836

Discriminator used for CWoLa

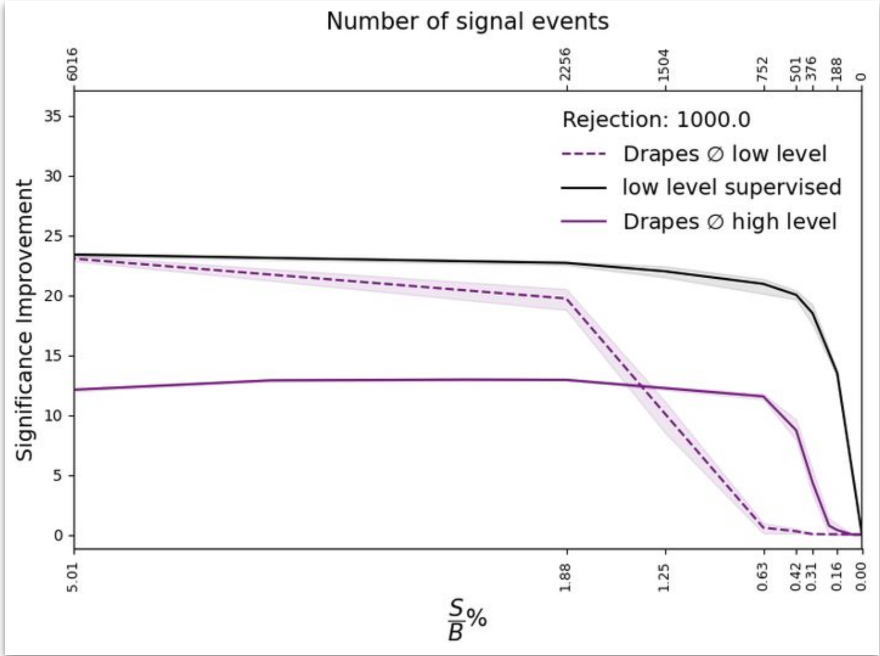
- The two jets are processed by the same network.
- The outputs are added and passed through MLP.



Drapes for constituent level

Huge improvement in SIC for several dopings.

High level features still performant for lower signal strengths!



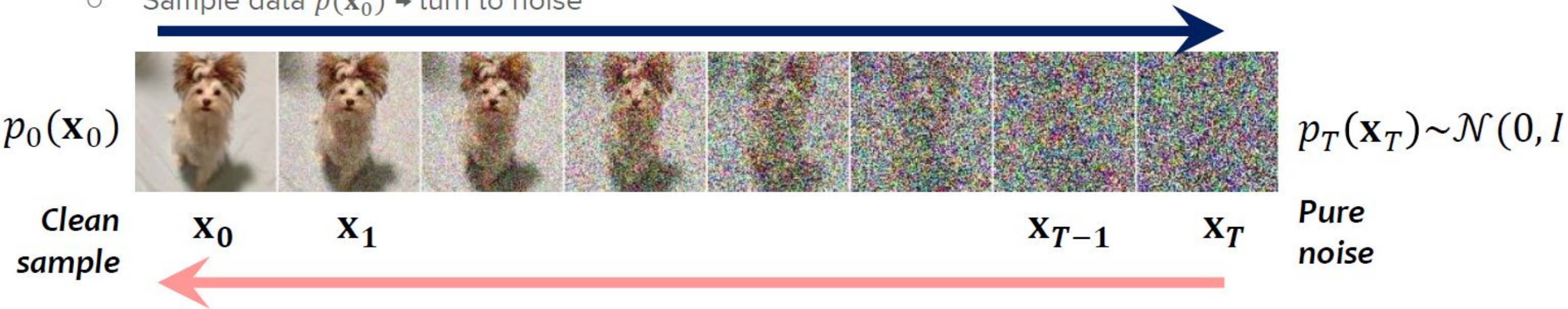
Conclusion

1. Diffusion perfectly viable for template generation
2. Partial diffusion saves time on template generation
3. Weakly supervised searches with low level data!

Backup

Diffusion Models

- **Forward / noising process**
  - Sample data $p(\mathbf{x}_0)$ → turn to noise

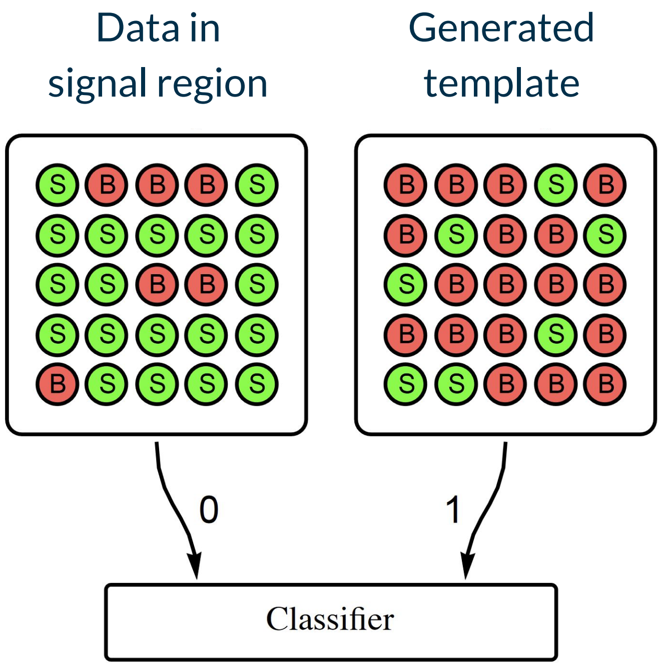


- **Reverse / denoising process**
  - Sample noise $p_T(\mathbf{x}_T)$ → turn into data

Established Task

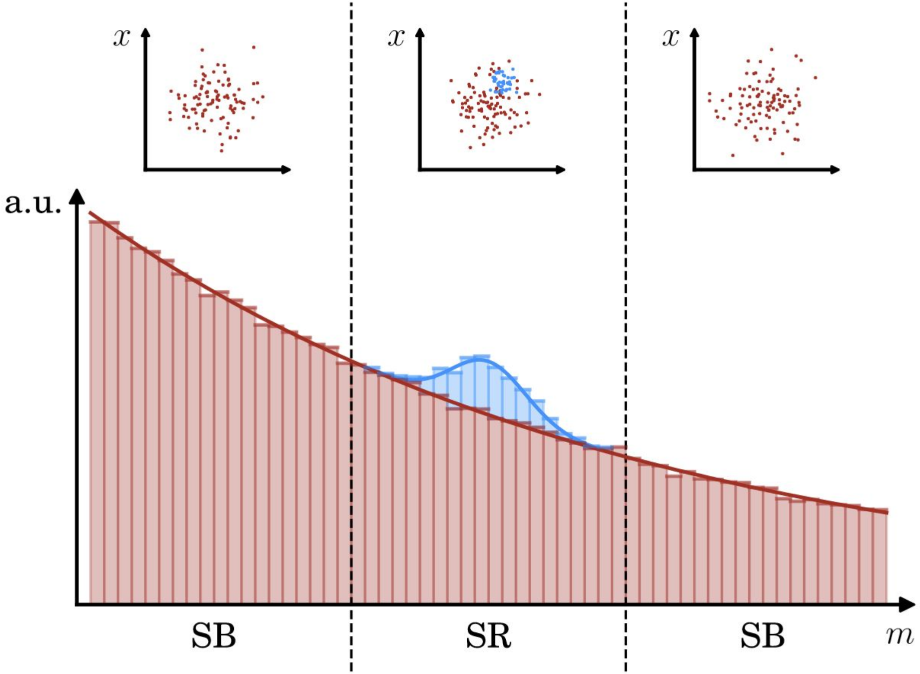
- Use CWOLA to look for anomalous samples

- Signal region contains: **B+S**
- Template should contain: **B (+S’)**



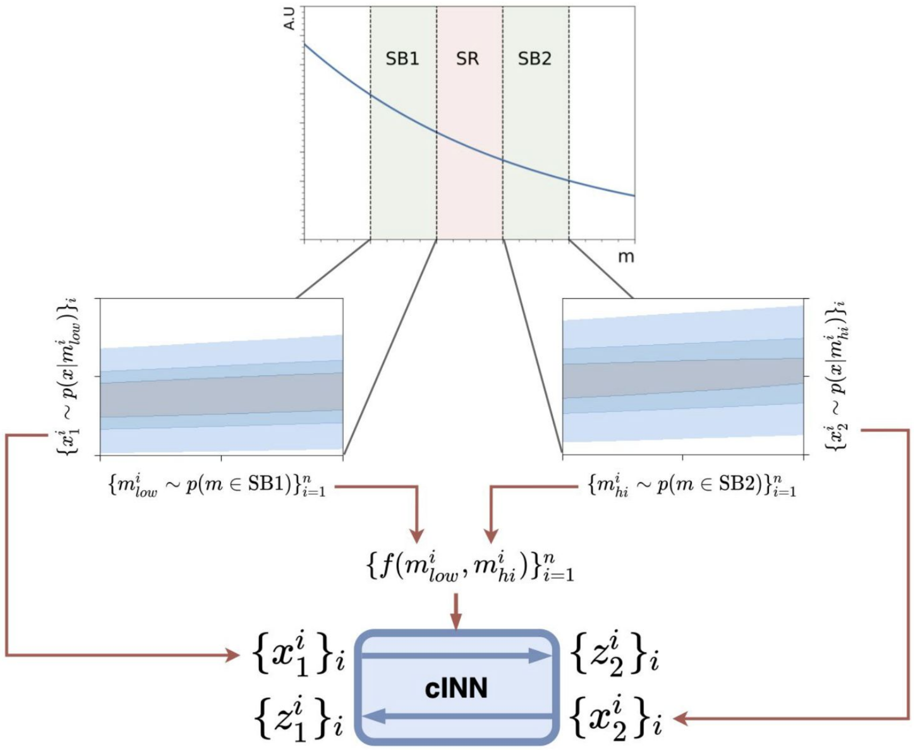
CATHODE

- Use **NORMALISING FLOW**
    - Train on sidebands
    - Condition on mass
    - Use to generate in signal region



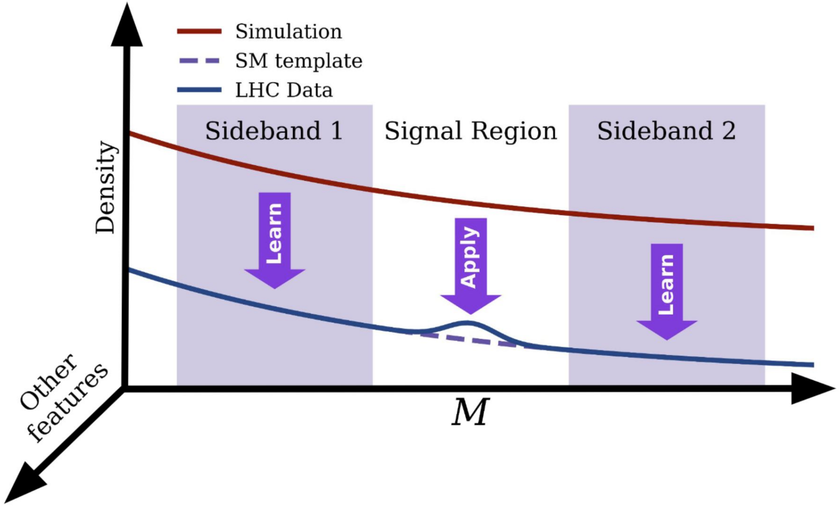
CURTAINS

- Rather than generate from scratch
- Learn how to **modify** data
    - **ie: Take a sample, give it a new mass, and morph**



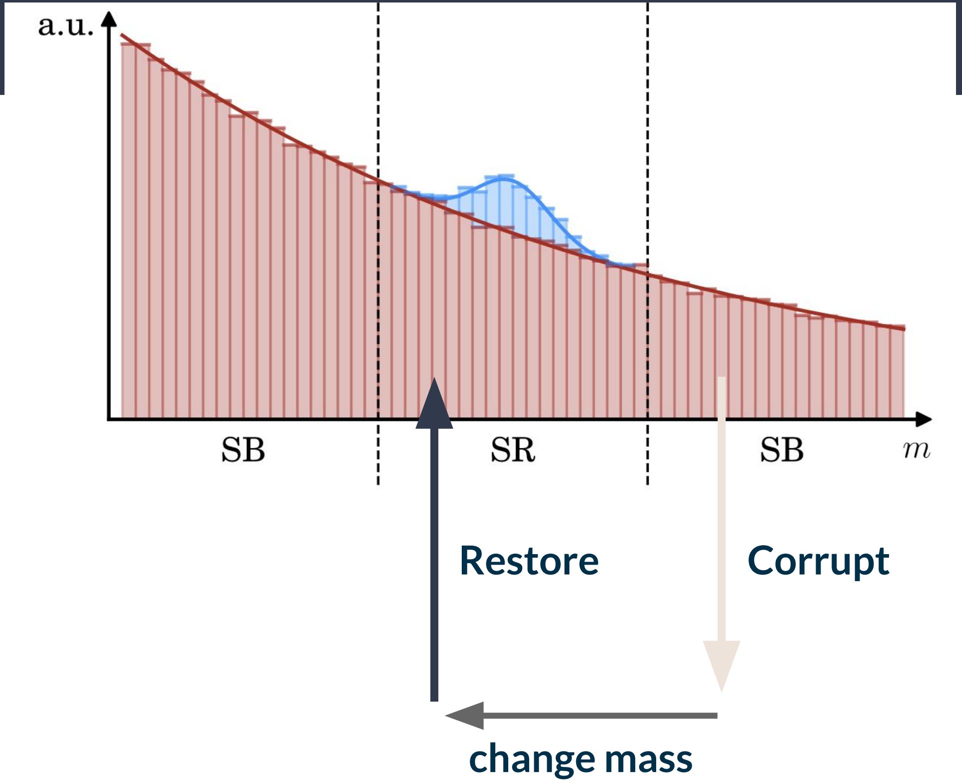
FETA

- Learn to transform **MC** to **DATA**
- Train by transforming sidebands
- Apply in signal region
- Learn how to **modify** data
    - Give it **new origin**



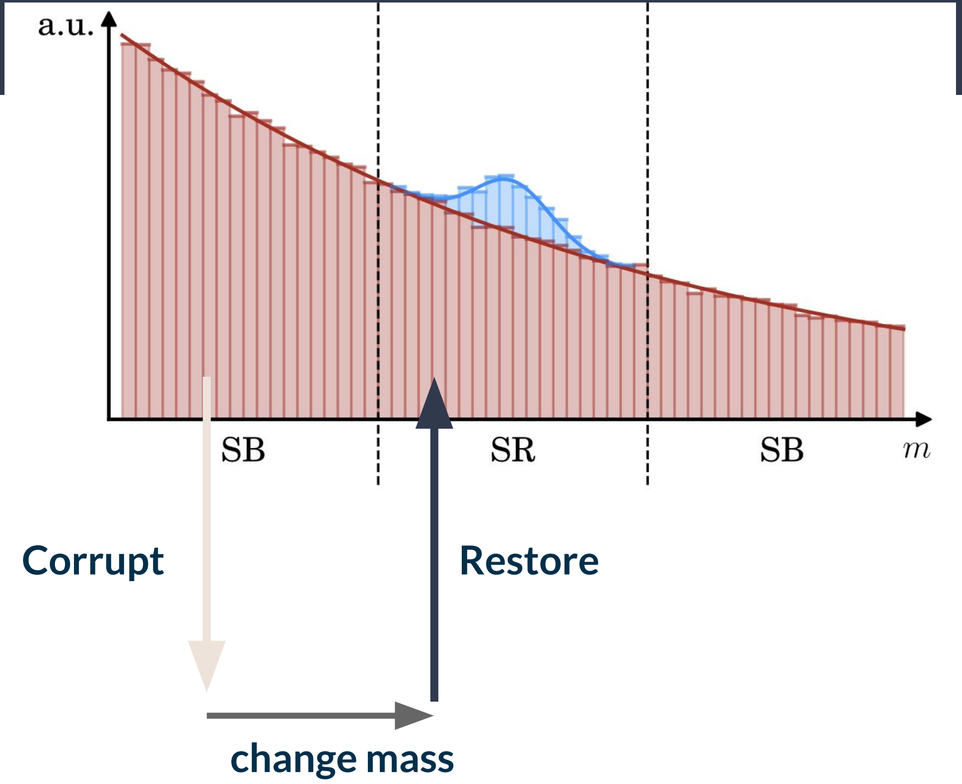
Drapes

- IMG2IMG allows us to **modify data**
- Where do we modify **our data from**?
    - **DRAPES SB**
    - From the sideband
    - Give sample new mass
    - CURTAINS



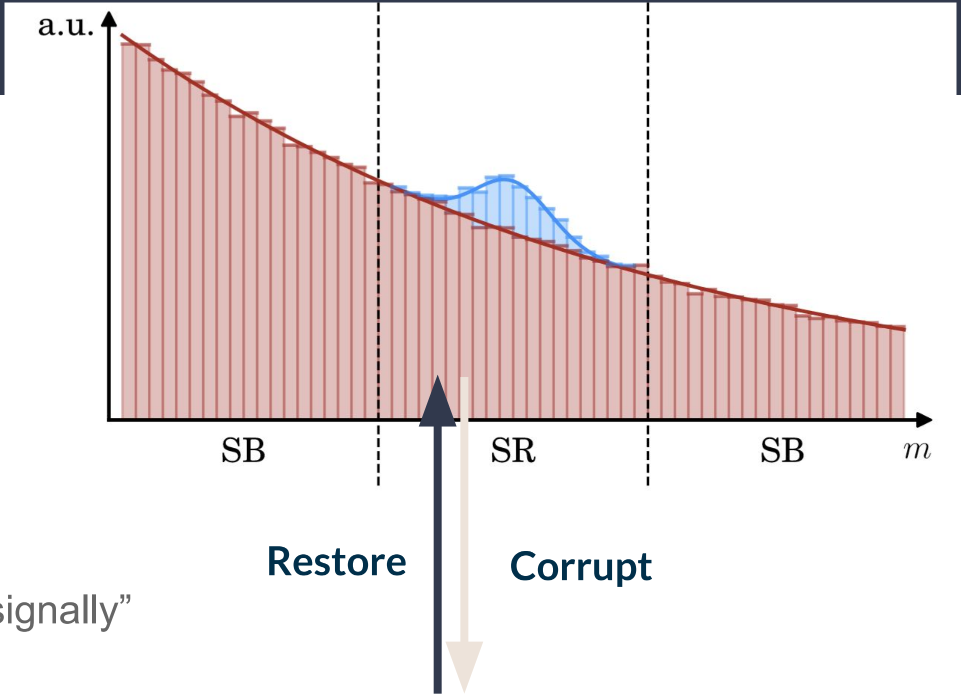
Drapes

- IMG2IMG allows us to **modify data**
- Where do we modify **our data from**?
  - **DRAPES SB**
  - From the sideband
  - Give sample new mass
  - CURTAINS



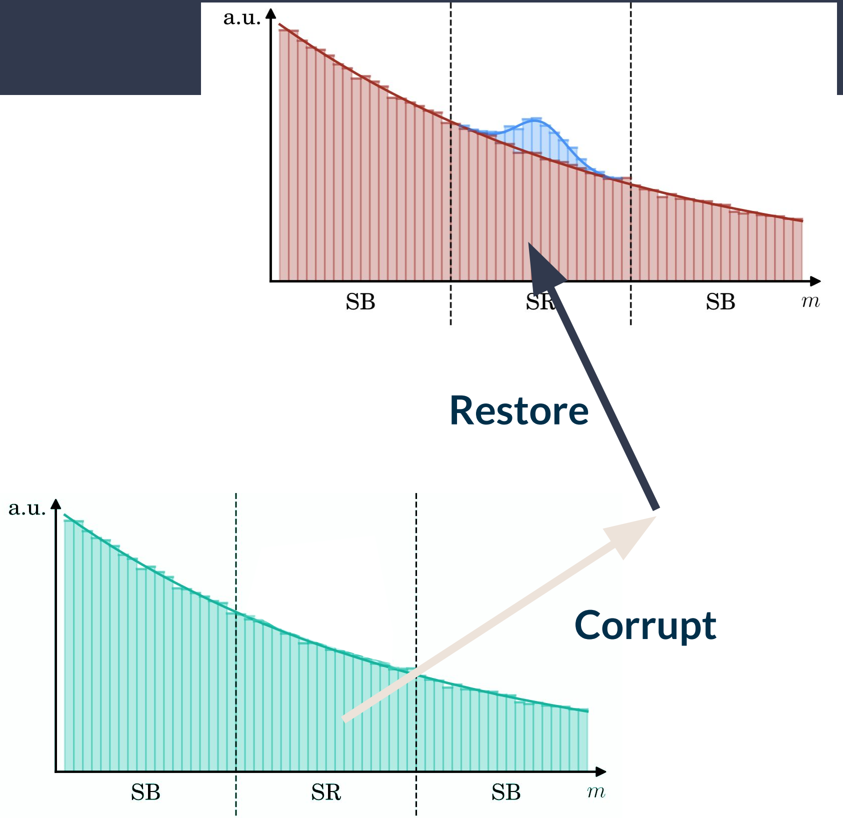
Drapes

- IMG2IMG allows us to **modify data**
- Where do we modify **our data from**?
    - **DRAPES SR**
    - From the signal region
    - Should make signal samples less “signally”



Drapes

- IMG2IMG allows us to **modify data**
- Where do we modify **our data from**?
    - <u>**DRAPES MC**</u>
    - From the another MC template
    - Change sample generation
    - FETA



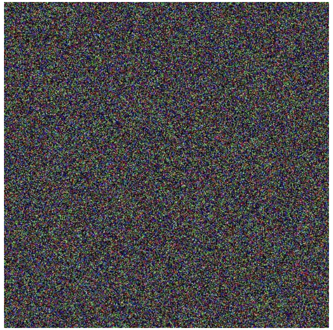
But there's more!

- Directly generating from noise is not the only way diffusion models can be used

IMG2IMG

t=0

t=strength



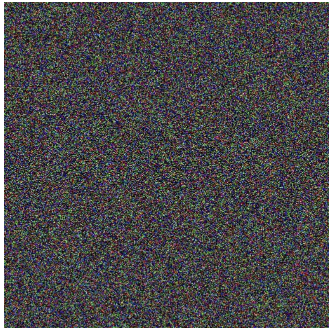
t=1

IMG2IMG

t=0

“picasso painting”

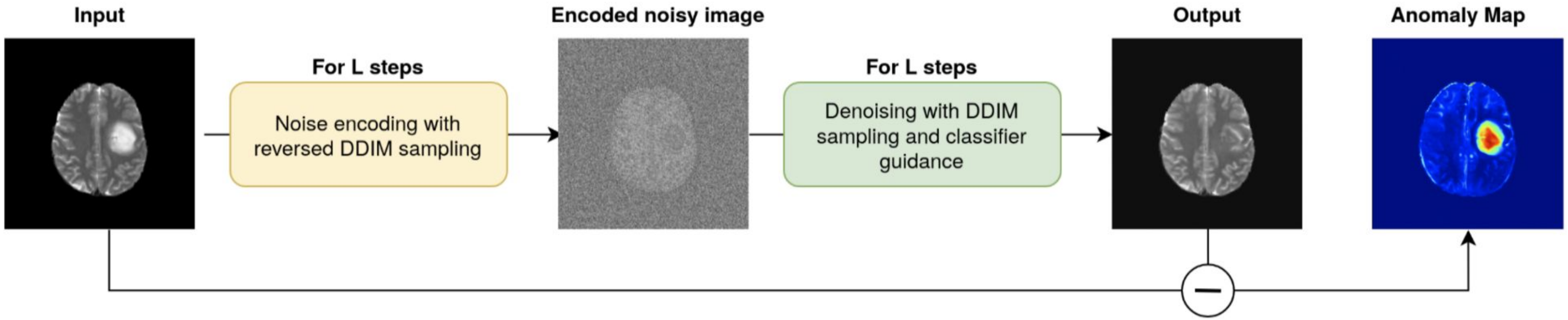
t=strength



t=1

Diffusion Anomaly Detection

- Method has seen success in image applications
- Won't be exactly how we will use it

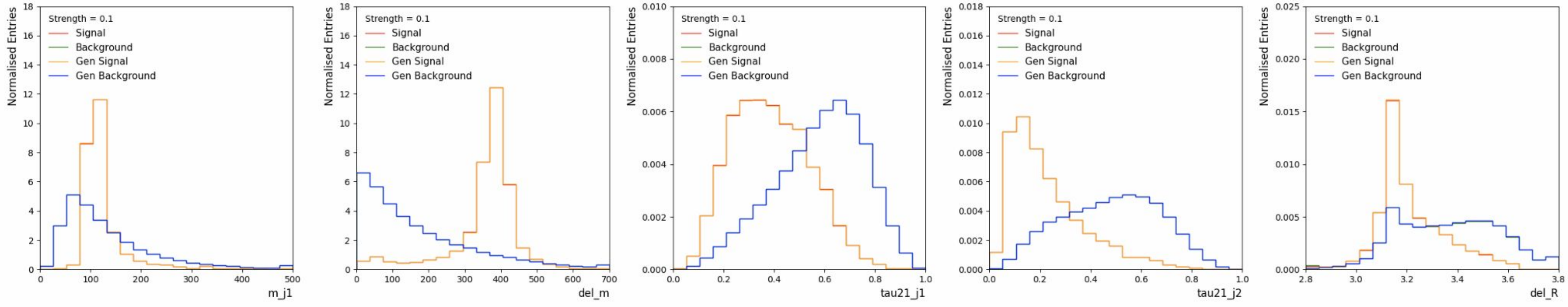


https://arxiv.org/pdf/2203.04306.pdf

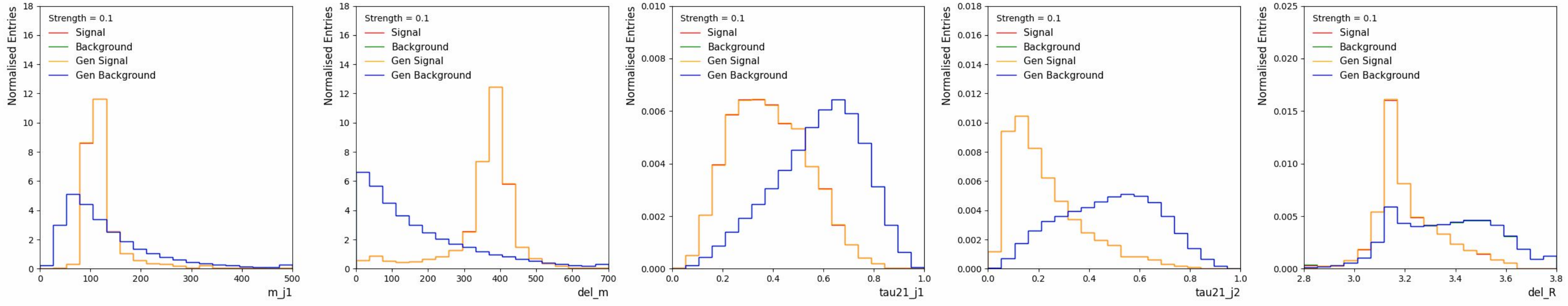
Diffusion Anomaly Detection

- We can **NOT** apply this type of anomaly detection in our data
- Images: **High** dimension, anomaly is **localised**, takes sample **off manifold**
- LHCO: **Low** dimension (5), anomaly is in the **over/under density** of a region
- So for now we stick to building the background templates + CWOLA

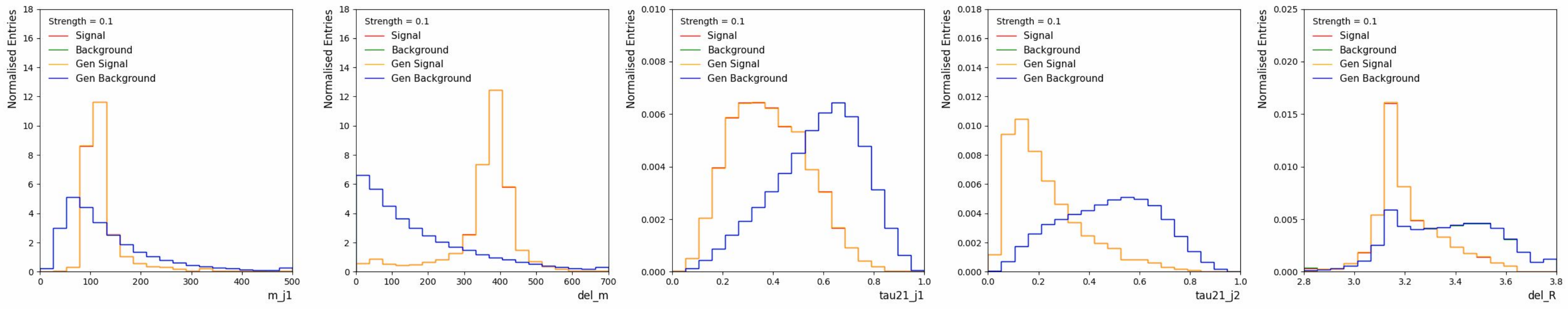
Drapes SR – Effect on Distributions



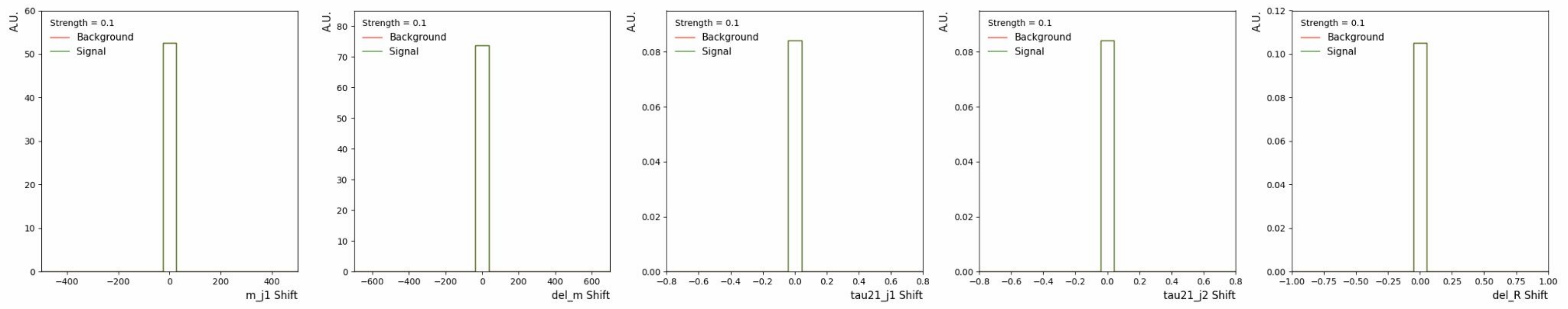
Drapes SR – Effect on Distributions



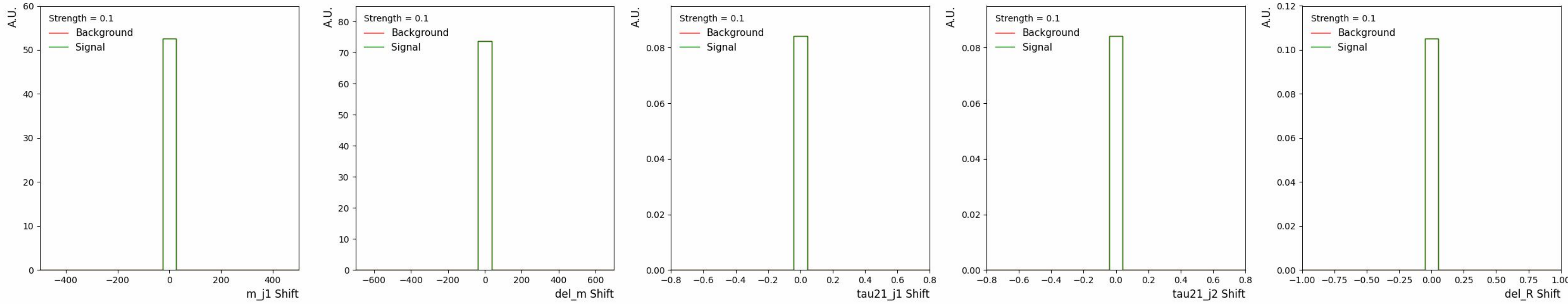
Drapes SR – Effect on Distributions



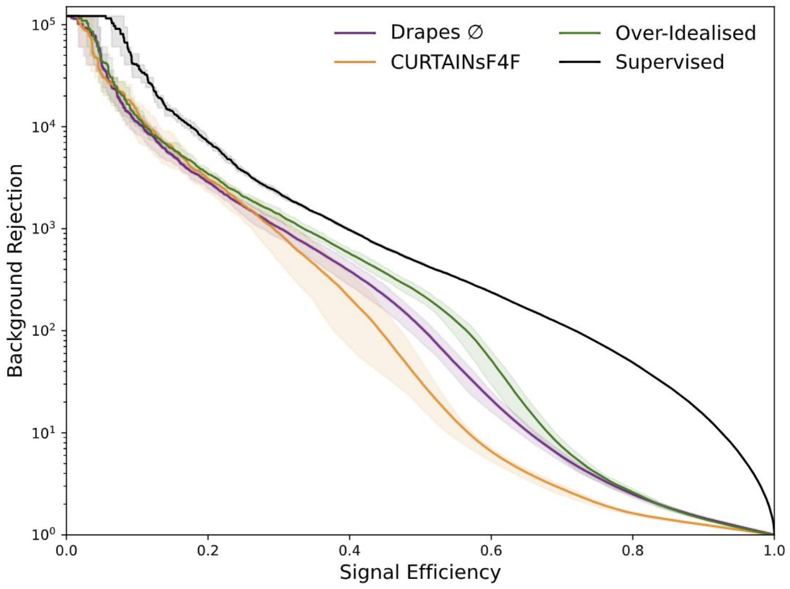
Drapes SR – Effect on Sample

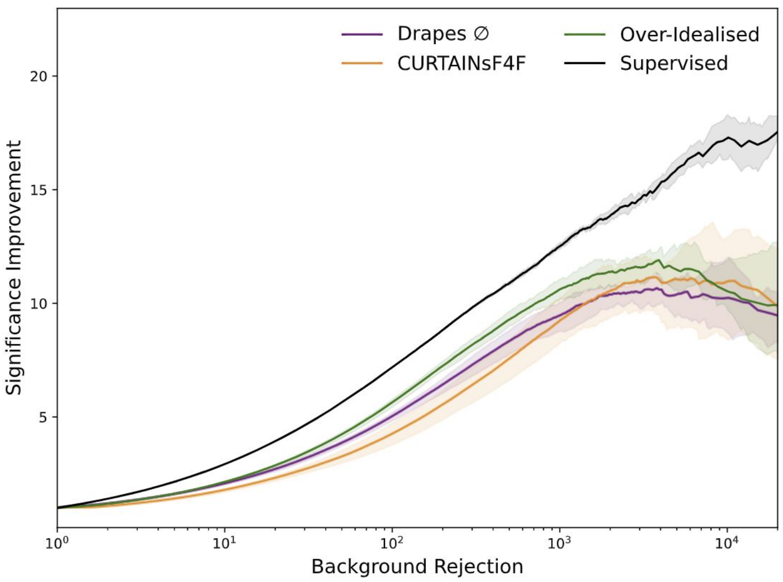


Drapes SR – Effect on Sample



Performance





Data doped with 1000 signal like events.